Chapter 2
A Multi-Feature Based Automatic Approach to Geospatial Record Linking

Ying Zhang
North China Electric Power University, China

Cheng Wang
North China Electric Power University, China

Puhai Yang
North China Electric Power University, China

Hui He
North China Electric Power University, China

Chaopeng Li
North China Electric Power University, China

Xiang Hu
North China Electric Power University, China

Gengrui Zhang
North China Electric Power University, China

Zhitao Guan
North China Electric Power University, China

ABSTRACT

This article describes how geographic information systems (GISs) can enable, enrich and enhance geospatial applications and services. Accurate calculation of the similarity among geospatial entities that belong to different data sources is of great importance for geospatial data linking. At present, most research works use the name or category of the entity to measure the similarity of geographic information. Although the geospatial relationship is significant for geographic similarity measure, it has been ignored by most of the previous works. This article introduces the geospatial relationship and topology, and proposes an approach to compute the geospatial record similarity based on multiple features including the geospatial relationships, category and name tags. In order to improve the flexibility and operability, supervised machine learning such as SVM is used for the task of classifying pairs of mapping records. The authors test their approach using three sources, namely, OpenStreetMap, Google and Wikimapia. The results showed that the proposed approach obtained high correlation with the human judgements.

DOI: 10.4018/978-1-5225-8054-6.ch002

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

1. INTRODUCTION

Nowadays, Geospatial Information System (GIS) has been extensively used to support all kinds of physical science and social science studies. However, there are a series of problems about the inconsistency, redundancy, ambiguity, and conflict of data in the collection of data for GIS project from diverse sources. Geographic information fusion focuses on the problem of combining geographic information from disparate sources so that accurate data can be retained, redundancies can be eliminated, and data conflicts can be reconciled (Samal et al., 2004; Feng, 2013). During the process of the information fusion, there must be variations of problems and no single measure to solve such classic problems. A typical approach for integration is often called feature fusion, which refers to the problem of improving the features in one source by combining the features of entities from another source or the others. The paper pays attention to the feature fusion problem for urban regions or other feature-rich areas. The primary question in feature fusion is the determination of the consistency between features in diverse sources, which is the focus of this paper.

At a glance, the problem seems to be trivial for reliable geospatial sources. However, in fact, the diverse sources with different commercial or non-profit purposes maybe have different spatial and temporal dimension, precision, accuracy in attributes, and cost. All of these bring about differentials in the representation of features in diverse sources. Thus, a raw overlay of the sources would not automatically make clear the consistency.

There is an interesting phenomenon that humans can quickly and exactly distinguish the matched features in multiple sources, even though seeming pretty inconsistency. They treat this by the comparison of names and positions of entities in sources, and the relations among those, and their domain knowledge. Similarly, in this paper, a multi-feature based measure is proposed for entity matching or mapping.

In this article, we describe a multi-feature based method, including the geospatial relationships, categories and name tags. Although the geospatial relationship is a significant characteristic of geospatial similarity measure, it has been ignored by most previous works. In order to improve the flexibility and operability, supervised machine learning, such as SVM, is designed for the task of classifying pairs of mapping records as either matches or not. The rest of the paper is organized as follows. Section 2 illustrates a motivation example of data integration. Section 3 introduces the related works. Section 4 discusses the presented method to retrieve and link the geospatial data in detail. The experimental results are presented in Section 5. Finally, the paper is concluded in Section 6 with an outlook to future challenges and work in this area.

2. MOTIVATION EXAMPLES

Let's consider Wikimapia, OpenStreetMap and Google Places as the example sources. Wikimapia provides names, types, latitudes, longitudes and polygon outlines for place entities, while OpenStreetMap gives elevation and address information, such as state and county name, in addition to the place names, amenities, longitudes, latitudes and polygons. In contrast, Google Places presents vicinity besides place names, types, latitudes and longitudes. For the same place, the extracted name from Wikimapia might be totally different from that provided by OpenStreetMap and Google Places. Moreover, for the same place, the location information may differ. For example, one place named "William Jefferson Clinton Middle School" by OpenStreetMap has coordinate "POLYGON((-118.2781488 34.0168423,-118.2771296

19 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/a-multi-feature-based-automatic-approach-to-geospatial-record-linking/222891

Related Content

Inter and Intra Cities Smartness: A Survey on Location Problems and GIS Tools
Ghada A. El Khayatand Nada Ahmad Fashal (2017). *Handbook of Research on Geographic Information Systems Applications and Advancements (pp. 296-320).*
www.irma-international.org/chapter/inter-and-intra-cities-smartness/169993

The Effects of Adoption of 3D Printing Technology on the Operational Performance of the Companies of Cross Border Entrepreneurs: An Empirical Study
Muath Surakji, Hani H. Al-dmourand Rand H. Al-Dmour (2018). *International Journal of 3-D Information Modeling (pp. 28-48).*
www.irma-international.org/article/the-effects-of-adoption-of-3d-printing-technology-on-the-operational-performance-of-the-companies-of-cross-border-entrepreneurs/238826

Distributed Geospatial Data Management for Entomological and Epidemiological Studies
Hugo Martinsand Jorge G. Rocha (2013). *Geographic Information Systems: Concepts, Methodologies, Tools, and Applications (pp. 1773-1793).*
www.irma-international.org/chapter/distributed-geospatial-data-management-entomological/70534

A Paradigm of Improving Land Information Management
Moha El-Ayachi (2019). *Geospatial Intelligence: Concepts, Methodologies, Tools, and Applications (pp. 1300-1319).*
www.irma-international.org/chapter/a-paradigm-of-improving-land-information-management/222948

Racial/Ethnic Diversity and Economy - A Broad Overview of U.S. Counties, 2000-2014: County Scale Diversity and Economy
Madhuri Sharma (2020). *International Journal of Applied Geospatial Research (pp. 18-41).*
www.irma-international.org/article/racialethnic-diversity-and-economy---a-broad-overview-of-us-counties-2000-2014/246007